
CSIT
CENTRE
FOR SECURE
INFORMATION
TECHNOLOGIES
DEEPSECURITY: APPLYING
DEEP LEARNING TO
HARDWARE SECURITY
QUEEN'S
UNIVERSITY
BELFAST
ESTP1845

DeepSecurity: Applying Deep Learning to Hardware Security

Overall Goal

To investigate the use of Deep Learning for security verification in EDA tools, specifically in relation to *Hardware Trojan detection* and *Side channel analysis* to allow non-security experts to receive feedback on how to improve the security of their designs prior to fabrication.

Rambus Cryptography Research

riscure

BAE SYSTEMS

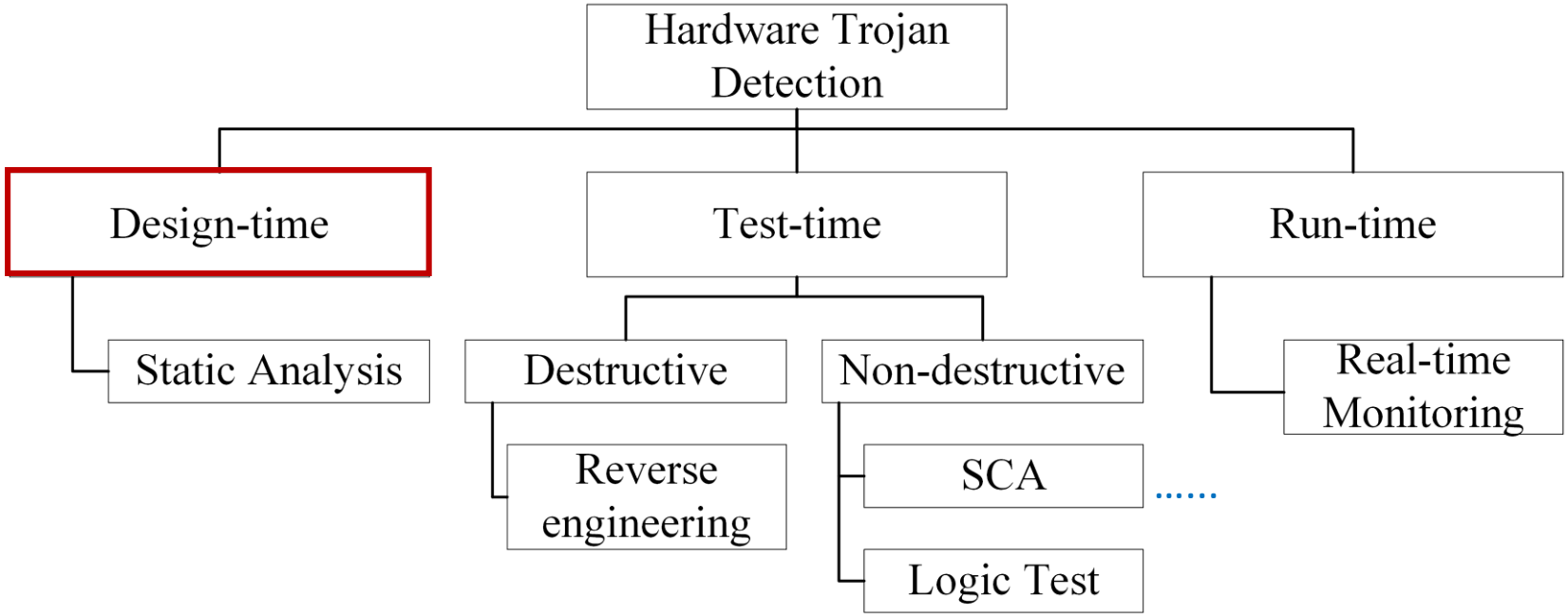
CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Hardware Trojan Detection

QUEEN'S UNIVERSITY BELFAST

Detecting Hardware Trojans – IC Production Stage

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES



Plan And Progress

Paper:
"An Improved Automatic Hardware Trojan Generation Platform" ISVLSI 2019, July 2019
" A Novel Feature Extraction Strategy for Hardware Trojan Detection " ISCAS 2020, (Submitted)

- ❑ Highly configurable in HT types;
- ❑ Triggered under rare conditions;
- ❑ HT-infected circuits with reports

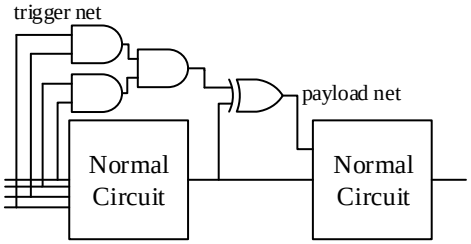
- ❑ Netlist parsing;
- ❑ Netlist block (in *N* logic levels);
- ❑ Trojan's structural features;
- ❑ Dynamic features (Switching activity)

- ❑ Unsupervised machine learning (K-means Clustering);
- ❑ Supervised machine learning (SVM);
- ❑ Deep Learning (under development)

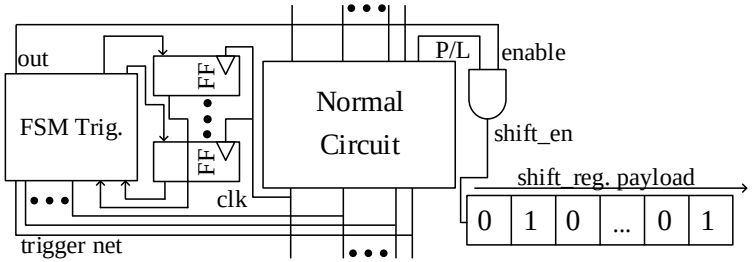
*SVM: Support Vector Machine

An Improved Automatic Hardware Trojan Generation Platform

Generated HT benchmark samples:



(a) Combinational Trojan with functional error payload



(b) FSM based sequential Trojan with a SHIFT based leakage circuit

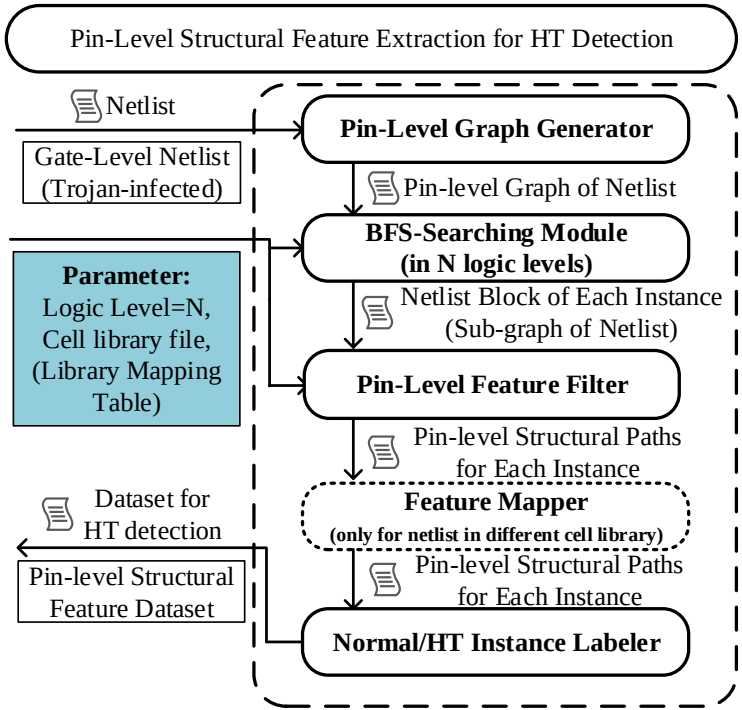
When compared with the COTD detection results from [1], who proposed a dynamic Hardware Trojan benchmark.

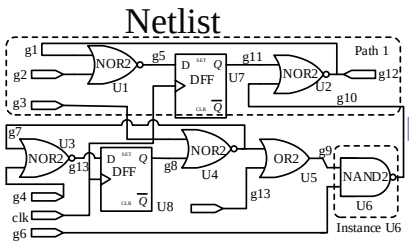
TABLE I
COMPARISON OF COTD-BASED HT DETECTION RESULTS

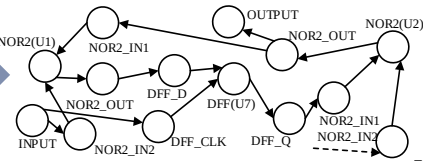
Benchmarks	Trigger Conditon (Rare/Total)	Type	HTs in [1]		Generated HTs [2]	
			FPR(%)	FNR(%)	FPR(%)	FNR(%)
s13207-c5_6	5/6	comb	25	0	0.39	23
s13207-s5_6	5/6	seq	0.11	0	0.41	19
s15850-c5_6	5/6	comb	27	0	12	25
s15850-s5_6	5/6	seq	0.09	0	0.11	17
s35932-c5_6	5/6	comb	60	0	15	33
s35932-s5_6	5/6	seq	0.08	0	0.03	20

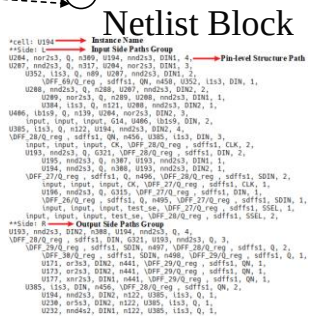
[1] J. Cruz, Y. Huang, P. Mishra, and S. Bhunia, "An automated configurable trojan insertion framework for dynamic trust benchmarks," in Proc. Design, Automation Test in Europe Conf. Exhibition, March 2018

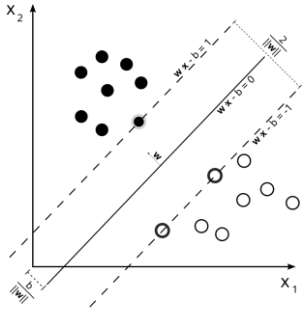
A Novel Feature Extraction Strategy for Hardware Trojan Detection







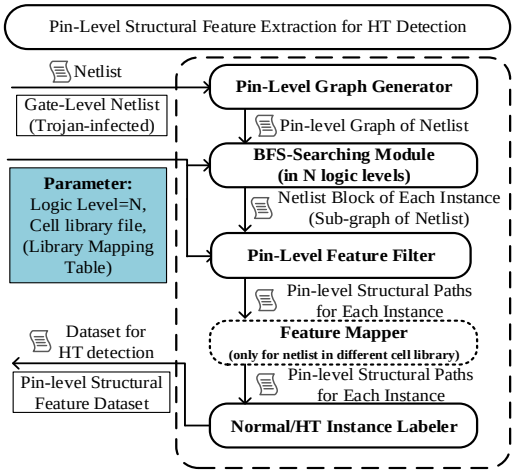




Next Steps

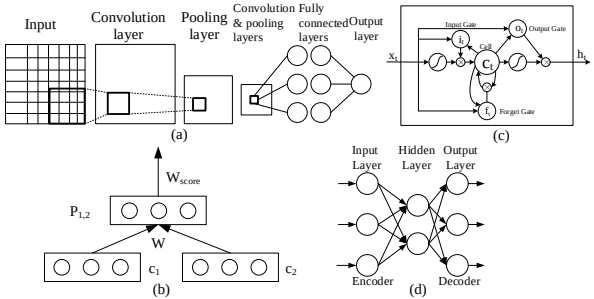
- Deep Learning-based HT detection model
- Testing:
  - Training Set: HT-infected benchmarks generated from our HT generation platform.
  - Testing Set: open-sourced HT benchmarks.

HT-infected Netlists



Novel Feature Extraction Program

HT Feature Traces



Deep Learning Model

Applying Deep Learning to Side-channel Analysis

QUEEN'S UNIVERSITY BELFAST

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Context and motivation

- Deep learning shows potential in improving side-channel analysis.
- Available deep learning models are not really designed for side-channel attacks.
- Evaluating DL-based SCA attacks and understanding the leakage that allows successful attacks can be used to improve physical attack countermeasures for cryptographic implementations.

ASCAD database

E. Prouff *et al.* Study of deep learning techniques for SCA and Introduction to ASCAD Database, Cryptology ePrint, Report 2018/053

- ASCAD database experiments
  - 2nd order masked AES-128 implementation on ATMega processor
  - Traces are synchronized and slightly de-synchronized
- ASCAD database with FIXED key
  - 50,000 traces for learning (50,000 traces for single key)
  - 10,000 traces for attacking, 700 samples per trace
- ASCAD database with VARIABLE key
  - 200,000 traces for learning (781 traces per key in average)
  - 100,000 traces for attacking, 1,400 samples per trace

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Models

- Attacking model: output value of 3rd Sbox at 1st round

$$Sbox_{out}[3] = Sbox(Plaintext[3] \oplus Key[3])$$

- Proposed Deep Learning models
  - Convolutional Neural Network
  - Considered general knowledge in Side-channel analysis
  - Target generic AES implementation (regardless of countermeasure used).
  - Different convolutional filter kernel sizes
- Reference models (in ASCAD database)
  - First order template model
  - First order Multi Layer Perceptron (MLP)
  - Multi Layer Perceptron with input batch normalization
  - VGG16 based CNN

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Assumption of attacker

- Attacker can profile plaintext and/or ciphertext
- Attacker can profile keys on the device
- Attacker does not know specific AES implementation details but understands that the designer may or may not have applied countermeasures e.g time shifting, 1st or high order masking, dual-rail logic, etc.

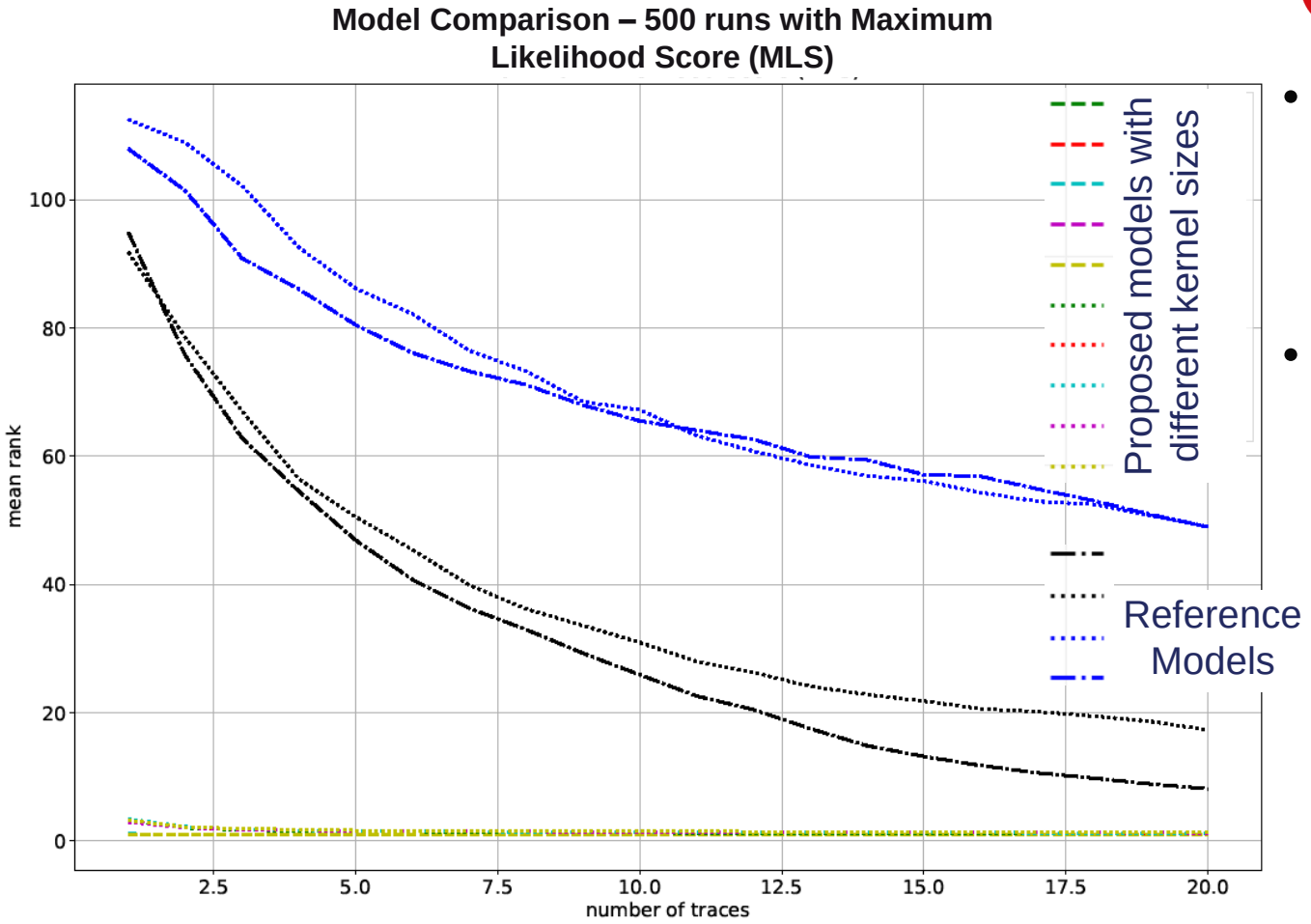
CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Model Development & Evaluation methods

- Use TensorFlow library for DL models

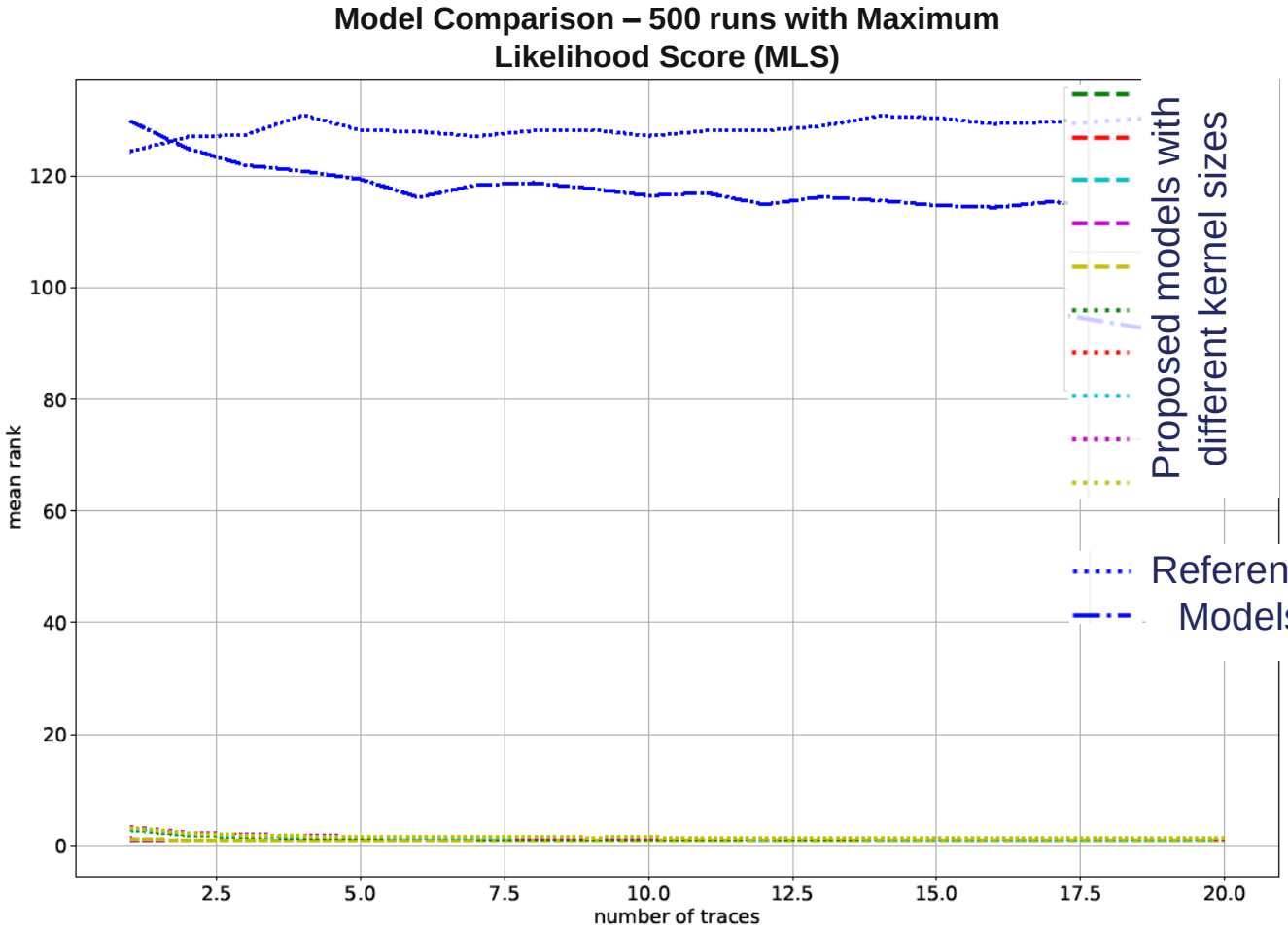
- Cross evaluation
  - Training on synchronized profile data and attacking on desynchronized attack data on the same database
  - Training on desynchronized profile data and attacking on synchronized attack data on the same database

Results for Fixed key, synchronized data



- Improves upon reference ML models (shown in blue & black)
- Our models are able to find correct sub key with a single trace

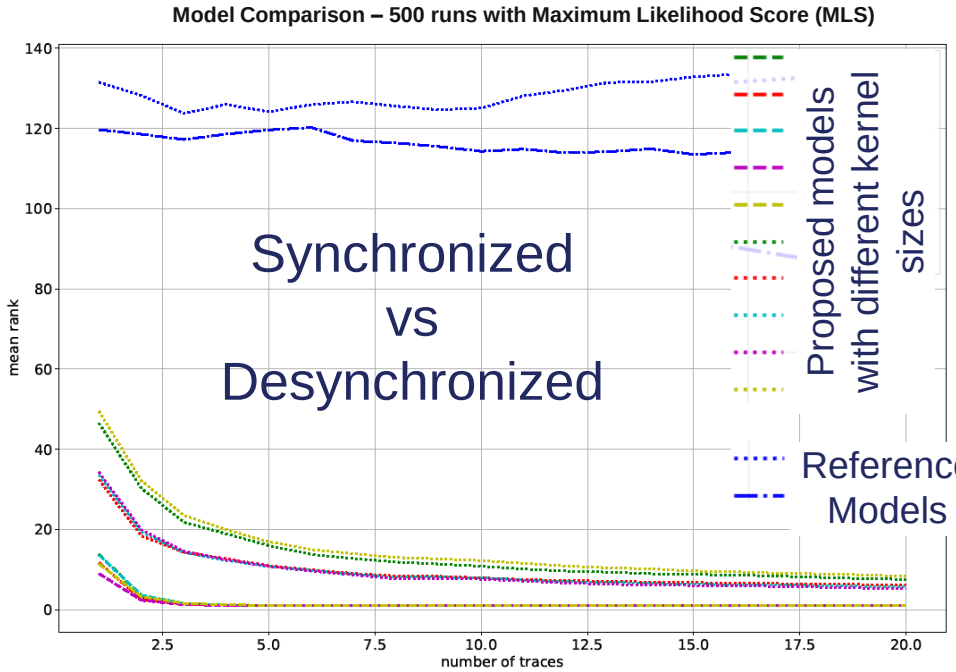
Results for Fixed key, desynchronized data

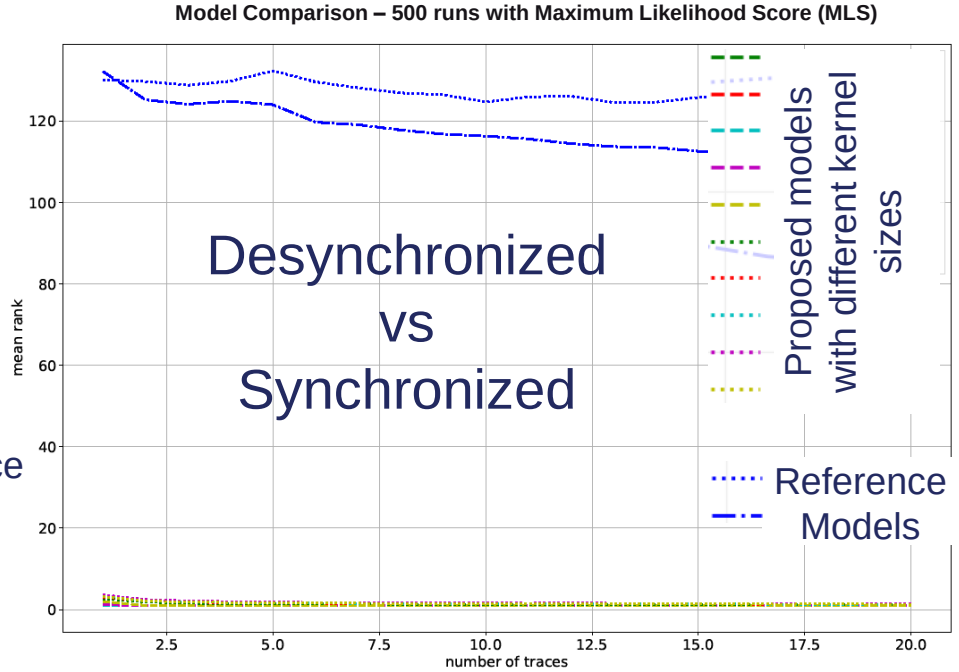


- Improves upon reference ML models (blue lines)
- All our models are able to find correct sub key within single traces

Results for Fixed key, cross evaluation

CSIT CENTRE FOR SECURE INFORMATION

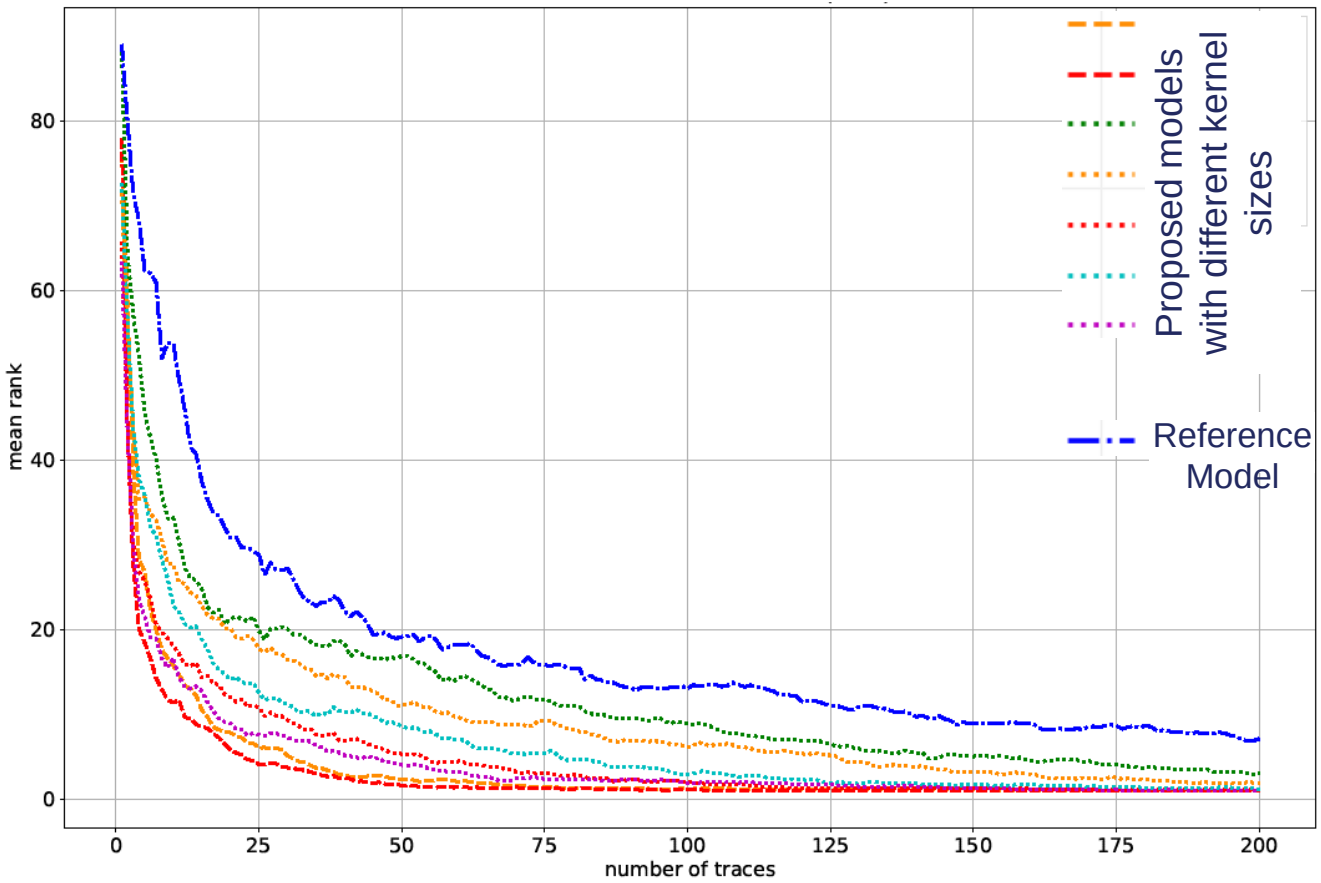




- Decrease in performance of models is observed when trained on synchronized data to attack desynchronized data
- No change observed when models trained on desynchronized data to attack synchronized data

Results for Variable key, synchronized data

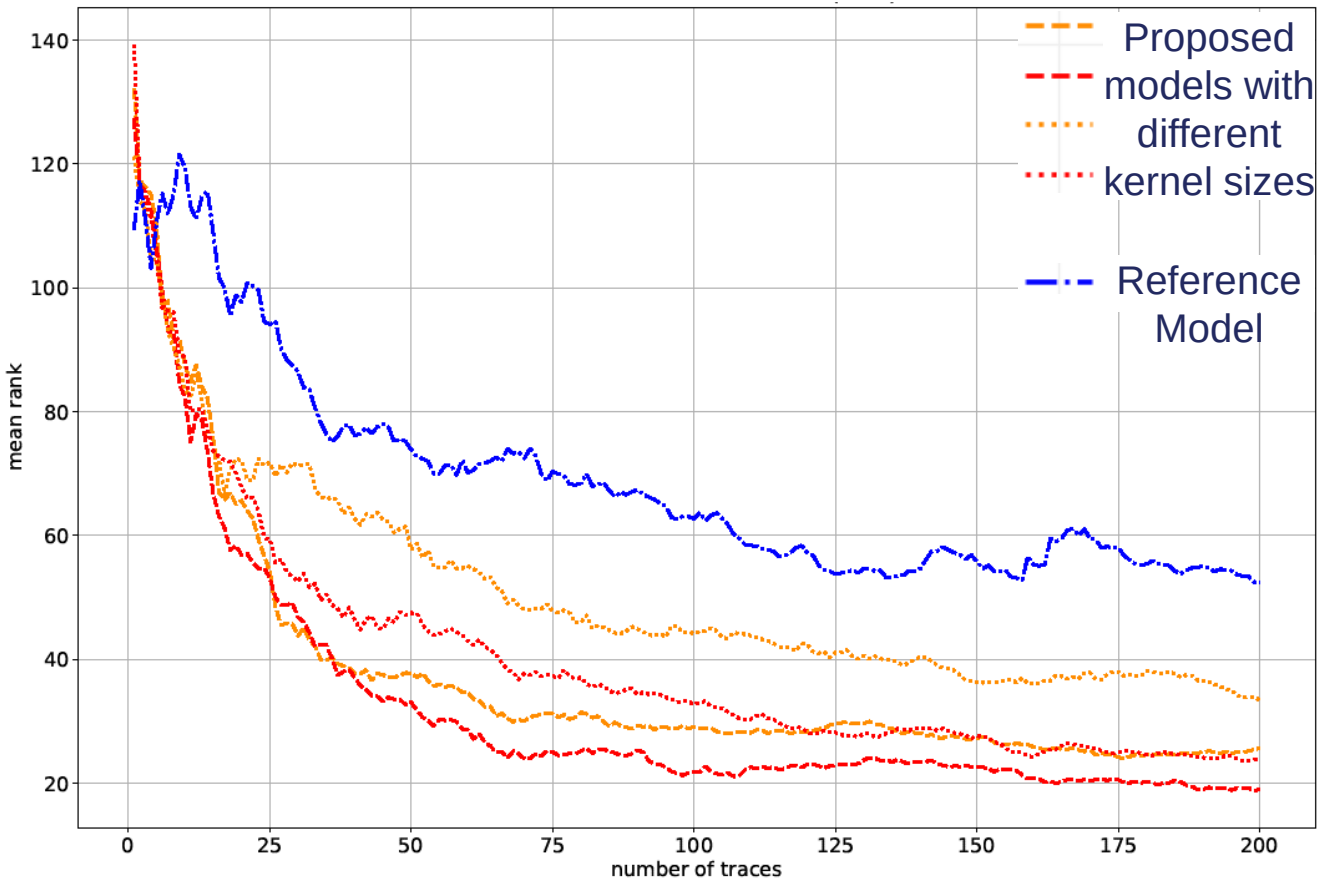
CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES



- Improves upon reference ML models (blue lines)
- Our CNN models are successful - achieve a key rank of 2 with 50 traces (2^{16} = 65,536 additional brute force attack required)

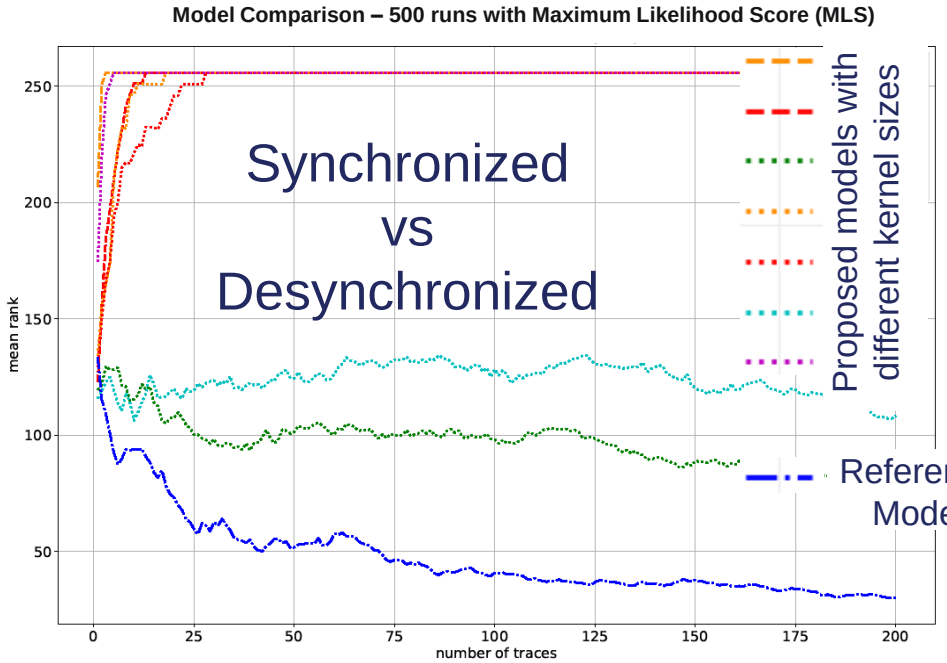
Results for Variable key, desynchronized data

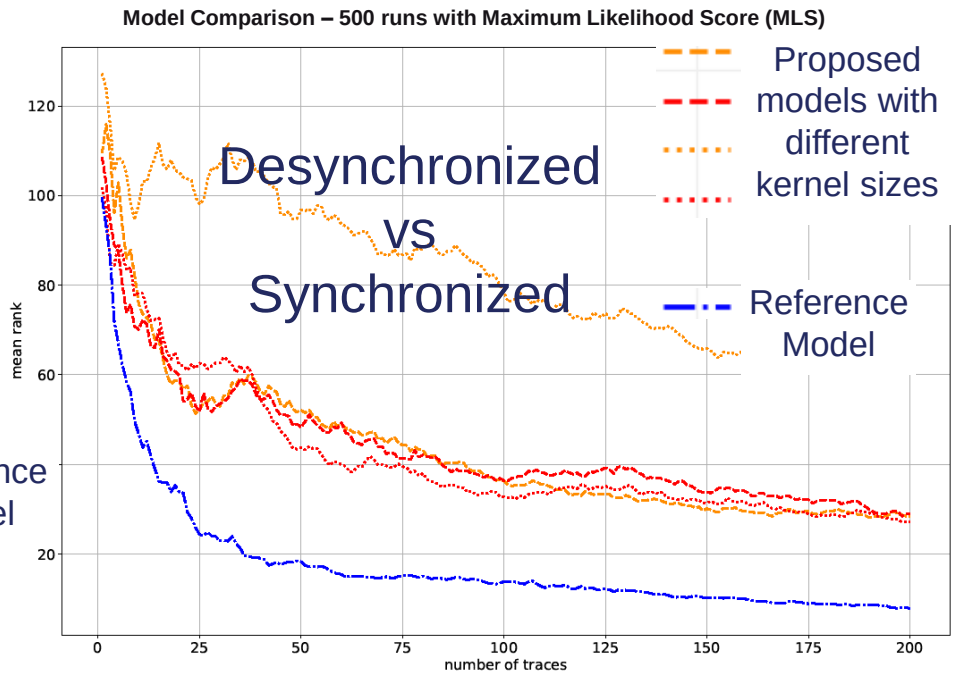
CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES



- Improves upon reference ML models (blue lines)
- Our CNN models achieve key rank 20 with 100 traces

Results for Variable key, cross evaluation





- Decrease in performance of proposed models observed in both situations
- May be due to small no. of traces available per sub key?

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Next steps

- Understand observed results in cross-evaluation attack for a variable key
- Application of proposed model to different databases, and different hardware based countermeasures, e.g. dual-rail logic approaches.
- Understand leakage that makes DL attacks possible in order to build stronger countermeasures.

Thank you